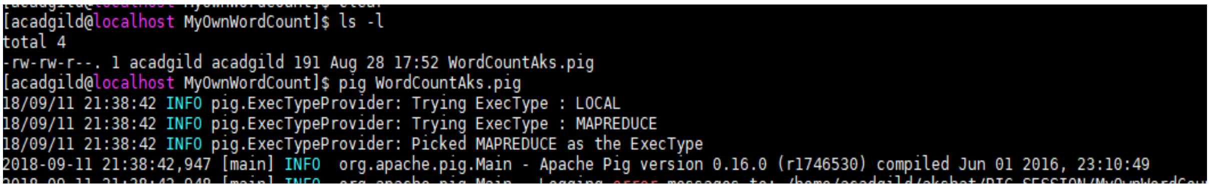
Assignment Submission- Session 7

Task 1: Write a program to implement wordcount using Pig.

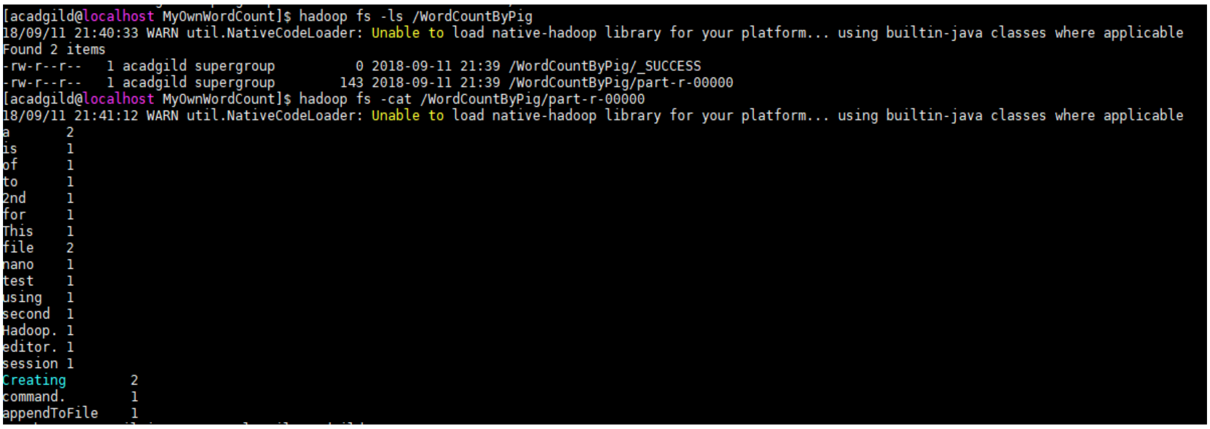
Solution:

Going to the directory where wordcount pig script is placed. The input file location needs to be given in the script. Running as below screenshot:



```
[acadgild@localhost MyOwnWordCount]$ ls -l
total 4
-rw-rw-r--. 1 acadgild acadgild 191 Aug 28 17:52 WordCountAks.pig
[acadgild@localhost MyOwnWordCount]$ pig WordCountAks.pig
18/09/11 21:38:42 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/09/11 21:38:42 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
18/09/11 21:38:42 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2018-09-11 21:38:42,947 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
```

The output generated is as screenshot:



```
[acadgild@localhost MyOwnWordCount]$ hadoop fs -ls /WordCountByPig
18/09/11 21:40:33 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup          0 2018-09-11 21:39 /WordCountByPig/_SUCCESS
-rw-r--r--   1 acadgild supergroup        143 2018-09-11 21:39 /WordCountByPig/part-r-00000
[acadgild@localhost MyOwnWordCount]$ hadoop fs -cat /WordCountByPig/part-r-00000
18/09/11 21:41:12 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
a       2
is      1
of      1
to      1
2nd     1
for     1
This    1
file    2
nano    1
test    1
using   1
second  1
Hadoop. 1
editor. 1
session 1
Creating        2
command.        1
appendToFile    1
```

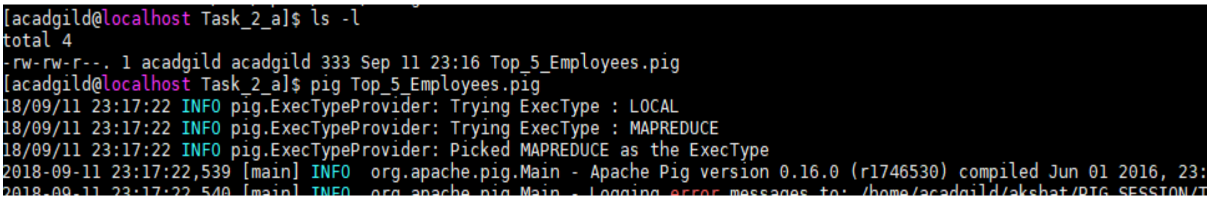
We have employee_details and employee_expenses files. Use local mode while running Pig and write Pig Latin script to get below results:

employee_details (EmpID,Name,Salary,EmployeeRating)

employee_expenses(EmpID,Expence)

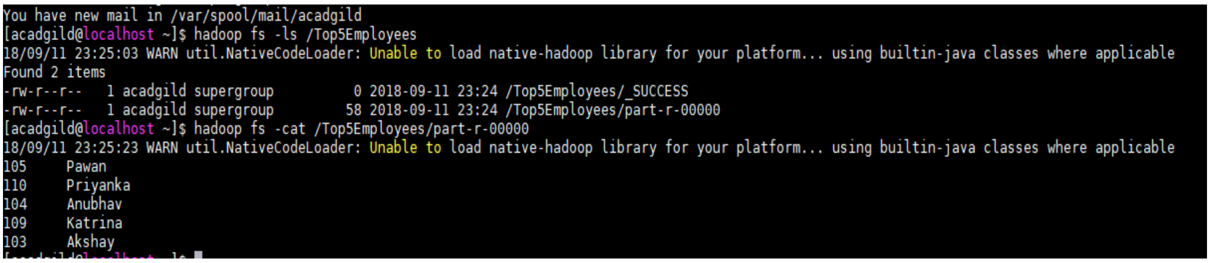
Task 2(a): Top 5 employees (employee id and employee name) with highest rating. (In case two employees have same rating, employee with name coming first in dictionary should get preference).

Solution: Going to the directory where pig script is placed. The input file location needs to be given in the script. Running as below screenshot:



```
[acadgild@localhost Task_2_a]$ ls -l
total 4
-rw-rw-r--. 1 acadgild acadgild 333 Sep 11 23:16 Top_5_Employees.pig
[acadgild@localhost Task_2_a]$ pig Top_5_Employees.pig
18/09/11 23:17:22 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/09/11 23:17:22 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
18/09/11 23:17:22 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2018-09-11 23:17:22,539 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:
2018-09-11 23:17:22,540 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/acadgild/akshat/PIG_SESSION/T
```

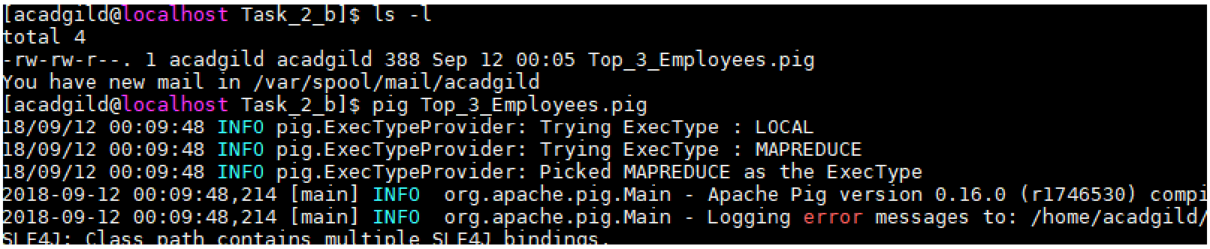
The output generated is as screenshot, also storing the output in a file (see script for more info):



```
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ hadoop fs -ls /Top5Employees
18/09/11 23:25:03 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup          0 2018-09-11 23:24 /Top5Employees/_SUCCESS
-rw-r--r--   1 acadgild supergroup         58 2018-09-11 23:24 /Top5Employees/part-r-00000
[acadgild@localhost ~]$ hadoop fs -cat /Top5Employees/part-r-00000
18/09/11 23:25:23 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
105     Pawan
110     Priyanka
104     Anubhav
109     Katrina
103     Akshay
[acadgild@localhost ~]$
```

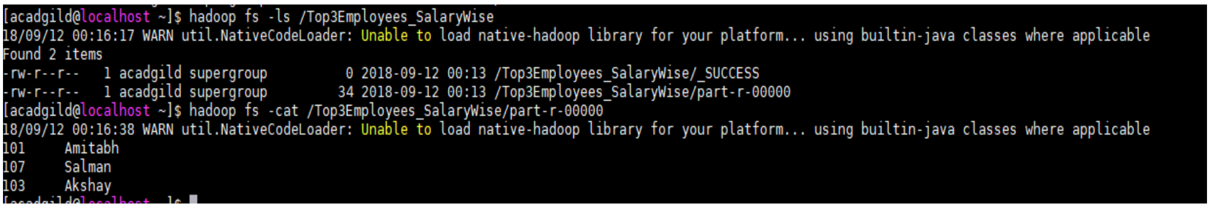
Task 2(b): Top 3 employees (employee id and employee name) with highest salary, whose employee id is an odd number. (In case two employees have same salary, employee with name coming first in dictionary should get preference)

Solution: Going to the directory where pig script is placed. The input file location needs to be given in the script. Running as below screenshot:



```
[acadgild@localhost Task_2_b]$ ls -l
total 4
-rw-rw-r--. 1 acadgild acadgild 388 Sep 12 00:05 Top_3_Employees.pig
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost Task_2_b]$ pig Top_3_Employees.pig
18/09/12 00:09:48 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/09/12 00:09:48 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
18/09/12 00:09:48 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2018-09-12 00:09:48,214 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compi
2018-09-12 00:09:48,214 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/acadgild/
SLF4J: Class path contains multiple SLF4J bindings.
```

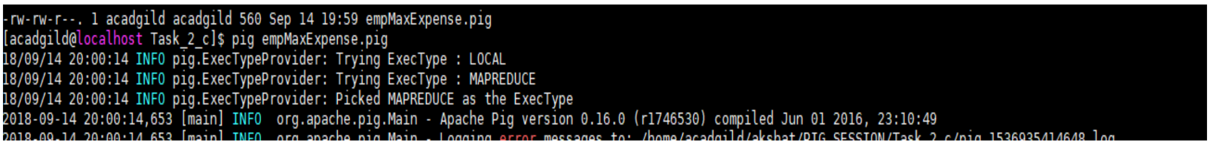
The output generated is as screenshot, also storing the output in a file (see script for more info):



```
[acadgild@localhost ~]$ hadoop fs -ls /Top3Employees_SalaryWise
18/09/12 00:16:17 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup          0 2018-09-12 00:13 /Top3Employees_SalaryWise/_SUCCESS
-rw-r--r--   1 acadgild supergroup         34 2018-09-12 00:13 /Top3Employees_SalaryWise/part-r-00000
[acadgild@localhost ~]$ hadoop fs -cat /Top3Employees_SalaryWise/part-r-00000
18/09/12 00:16:38 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
101     Amitabh
107     Salman
103     Akshay
[acadgild@localhost ~]$
```

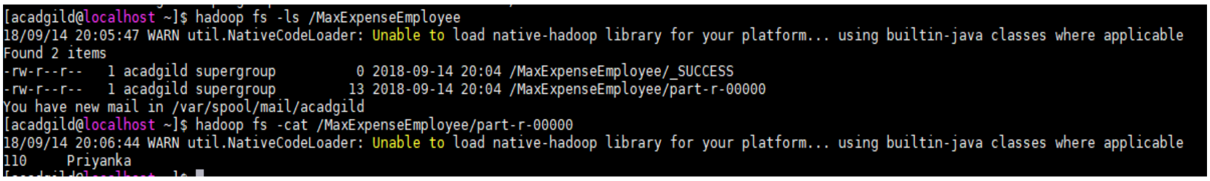
Task 2(c): Employee (employee id and employee name) with maximum expense (In case two employees have same expense, employee with name coming first in dictionary should get preference)

Solution: Going to the directory where pig script is placed. The input file location needs to be given in the script. Running as below screenshot:



```
-rw-rw-r--. 1 acadgild acadgild 560 Sep 14 19:59 empMaxExpense.pig
[acadgild@localhost Task_2_c]$ pig empMaxExpense.pig
18/09/14 20:00:14 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/09/14 20:00:14 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
18/09/14 20:00:14 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2018-09-14 20:00:14,653 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2018-09-14 20:00:14,653 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/acadgild/akshat/PIG_SESSION/Task_2_c/pig_1536935414648.log
```

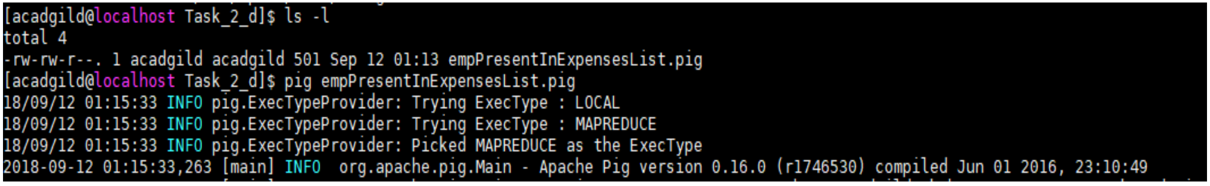
The output generated is as screenshot, also storing the output in a file (see script for more info):



```
[acadgild@localhost ~]$ hadoop fs -ls /MaxExpenseEmployee
18/09/14 20:05:47 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup          0 2018-09-14 20:04 /MaxExpenseEmployee/_SUCCESS
-rw-r--r--   1 acadgild supergroup         13 2018-09-14 20:04 /MaxExpenseEmployee/part-r-00000
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ hadoop fs -cat /MaxExpenseEmployee/part-r-00000
18/09/14 20:06:44 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
110     Priyanka
[acadgild@localhost ~]$
```

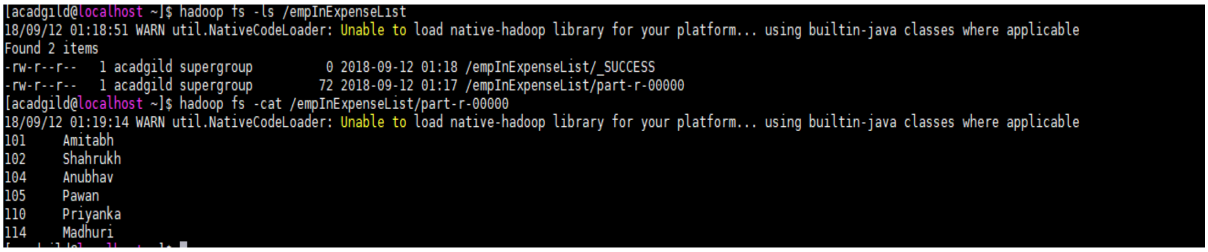
Task 2(d): List of employees (employee id and employee name) having entries in employee_expenses file.

Solution: Going to the directory where pig script is placed. The input file location needs to be given in the script. Running as below screenshot:



```
[acadgild@localhost Task_2_d]$ ls -l
total 4
-rw-rw-r--. 1 acadgild acadgild 501 Sep 12 01:13 empPresentInExpensesList.pig
[acadgild@localhost Task_2_d]$ pig empPresentInExpensesList.pig
18/09/12 01:15:33 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/09/12 01:15:33 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
18/09/12 01:15:33 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2018-09-12 01:15:33,263 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
```

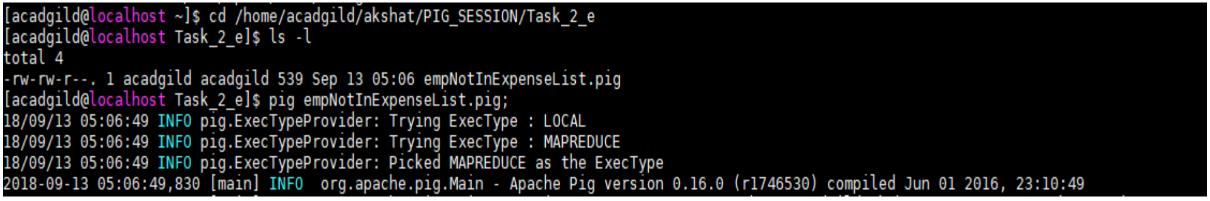
The output generated is as screenshot, also storing the output in a file (see script for more info):



```
[acadgild@localhost ~]$ hadoop fs -ls /empInExpenseList
18/09/12 01:18:51 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup          0 2018-09-12 01:18 /empInExpenseList/_SUCCESS
-rw-r--r--   1 acadgild supergroup         72 2018-09-12 01:17 /empInExpenseList/part-r-00000
[acadgild@localhost ~]$ hadoop fs -cat /empInExpenseList/part-r-00000
18/09/12 01:19:14 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
101     Amitabh
102     Shahrukh
104     Anubhav
105     Pawan
110     Priyanka
114     Madhuri
[acadgild@localhost ~]$
```

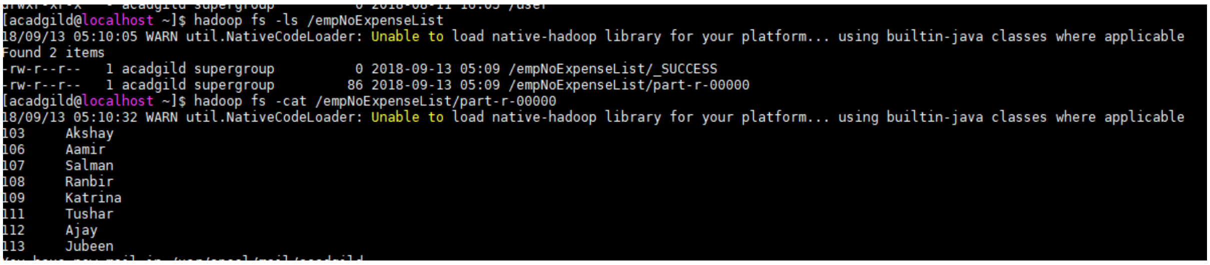
Task 2(e): List of employees (employee id and employee name) having no entry in employee_expenses file.

Solution: Going to the directory where pig script is placed. The input file location needs to be given in the script. Running as below screenshot:



```
[acadgild@localhost ~]$ cd /home/acadgild/akshat/PIG_SESSION/Task_2_e
[acadgild@localhost Task_2_e]$ ls -l
total 4
-rw-rw-r--. 1 acadgild acadgild 539 Sep 13 05:06 empNotInExpenseList.pig
[acadgild@localhost Task_2_e]$ pig empNotInExpenseList.pig;
18/09/13 05:06:49 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
18/09/13 05:06:49 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
18/09/13 05:06:49 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2018-09-13 05:06:49,830 [main] INFO  org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
```

The output generated is as screenshot, also storing the output in a file (see script for more info):



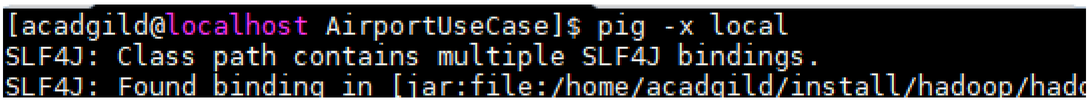
```
drwxr-xr-x   - acadgild supergroup          0 2018-08-11 10:05 /user
[acadgild@localhost ~]$ hadoop fs -ls /empNoExpenseList
18/09/13 05:10:05 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 acadgild supergroup          0 2018-09-13 05:09 /empNoExpenseList/_SUCCESS
-rw-r--r--   1 acadgild supergroup         86 2018-09-13 05:09 /empNoExpenseList/part-r-00000
[acadgild@localhost ~]$ hadoop fs -cat /empNoExpenseList/part-r-00000
18/09/13 05:10:32 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
103     Akshay
106     Aamir
107     Salman
108     Ranbir
109     Katrina
111     Tushar
112     Ajay
113     Jubeen
You have new mail in /var/spool/mail/acadgild
```

Task 3:

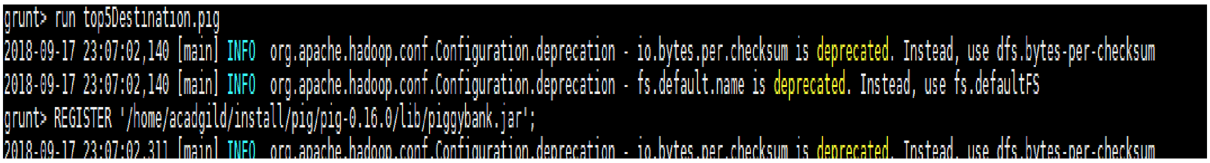
Implement the use case present in below blog link and share the complete steps along with screenshot(s) from your end.
https://acadgild.com/blog/aviation-data-analysis-using-apache-pig/

1. Find out the top 5 most visited destinations:

Running pig in local mode.

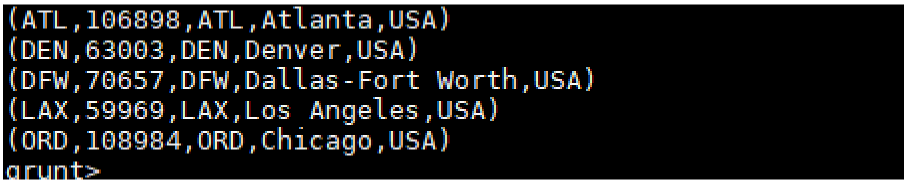


```
[acadgild@localhost AirportUseCase]$ pig -x local
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hado
```



```
grunt> run top5Destination.pig
2018-09-17 23:07:02,140 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-09-17 23:07:02,140 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> REGISTER '/home/acadgild/install/pig/pig-0.16.0/lib/piggybank.jar';
2018-09-17 23:07:02,311 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
```

Output:



```
(ATL,106898,ATL,Atlanta,USA)
(DEN,63003,DEN,Denver,USA)
(DFW,70657,DFW,Dallas-Fort Worth,USA)
(LAX,59969,LAX,Los Angeles,USA)
(ORD,108984,ORD,Chicago,USA)
grunt>
```

Script explanation:

→registering the piggybank jar to use the CSVExcelStorage class.
→loading the dataset using CSVExcelStorage.

→ using foreach getting year, flight number, origin and destination of flight.
→filtering out all rows that have destination as null.
→grouping filtered data by destination

→generating grouped column(destination) and number of times it has been destination.
→ordering in descending as per destination count

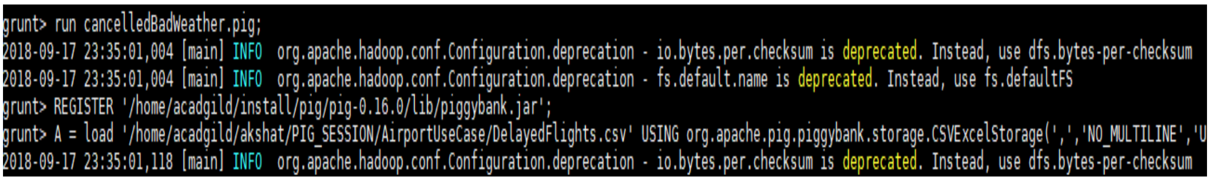
→Limiting result to top 5.

→ loading airport data to get the city, state and country info of top destinations.

→ using foreach getting city, state and country info of each airport/destination.

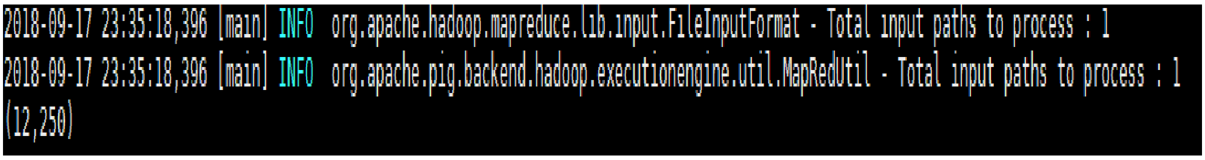
→joining both relation on common column destination.

2. Which month has seen the most number of cancellations due to bad weather?



```
grunt> run cancelledBadWeather.pig;
2018-09-17 23:35:01,004 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-09-17 23:35:01,004 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> REGISTER '/home/acadgild/install/pig/pig-0.16.0/lib/piggybank.jar';
grunt> A = load '/home/acadgild/akshat/PIG_SESSION/AirportUseCase/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','U
2018-09-17 23:35:01,118 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
```

Output:



```
2018-09-17 23:35:18,396 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2018-09-17 23:35:18,396 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(12,250)
```

Script explanation:

→registering the piggybank jar to use the CSVExcelStorage class.
→loading the dataset using CSVExcelStorage.

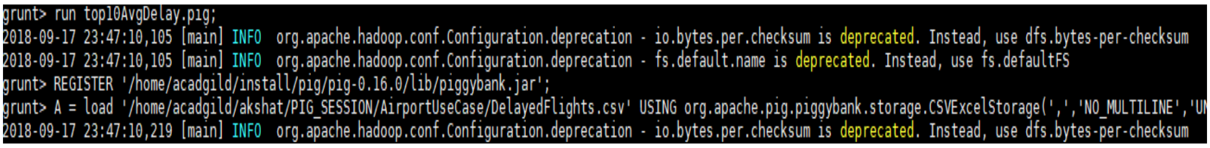
→ using foreach getting month, flight number, cancelled and cancellation code.
→filtering out all rows that have cancelled ==1 and cancellation code “B” for bad weather

→grouping filtered data by month.

→generating grouped column(month) and number of times of cancellation for that month.
→ordering in descending as per cancellation count

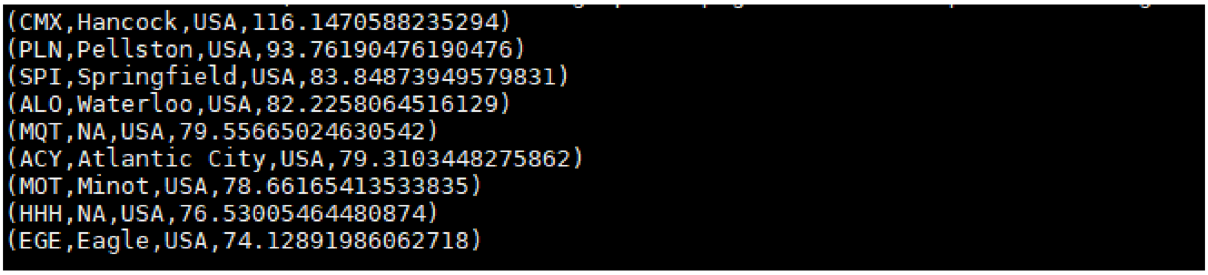
→Limiting result to top 1.

3. Top ten origins with the highest AVG departure delay.



```
grunt> run top10AvgDelay.pig;
2018-09-17 23:47:10,105 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-09-17 23:47:10,105 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> REGISTER '/home/acadgild/install/pig/pig-0.16.0/lib/piggybank.jar';
grunt> A = load '/home/acadgild/akshat/PIG_SESSION/AirportUseCase/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UN
2018-09-17 23:47:10,219 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
```

Output:



```
(CMX,Hancock,USA,116.1470588235294)
(PLN,Pellston,USA,93.76190476190476)
(SPI,Springfield,USA,83.84873949579831)
(ALO,Waterloo,USA,82.2258064516129)
(MQT,NA,USA,79.55665024630542)
(ACY,Atlantic City,USA,79.3103448275862)
(MOT,Minot,USA,78.66165413533835)
(HHH,NA,USA,76.53005464480874)
(EGE,Eagle,USA,74.12891986062718)
```

Script explanation:

→registering the piggybank jar to use the CSVExcelStorage class.
→loading the dataset using CSVExcelStorage.

→ using foreach getting origin and delay in departure.
→filtering out all rows that have origin OR departure is null.
→grouping filtered data by origin.

→generating grouped column(origin) and average dept delay.
→ordering in descending as per dept delay

→Limiting result to top 10.

→ loading airport data to get the origin, city and country info.

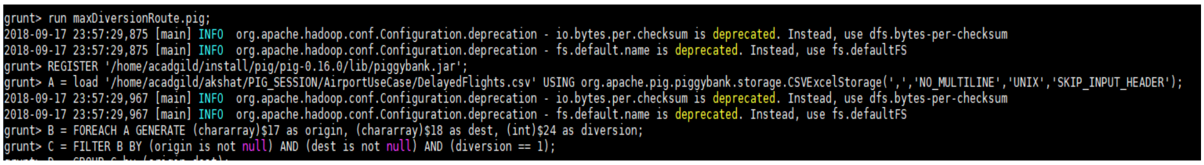
→ using foreach getting city, origin and country info of each airport/destination.

→joining both relation on common column origin.

→ getting required columns from joined relations.

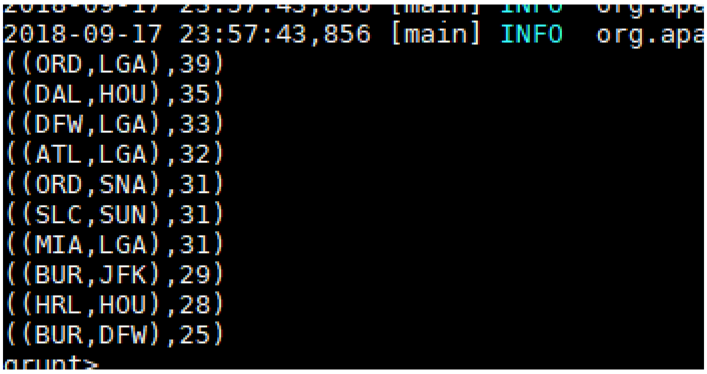
→ finally ordering again and dumping result.

4. Which route (origin & destination) has seen the maximum diversion?



```
grunt> run maxDiversionRoute.pig;
2018-09-17 23:57:29,875 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-09-17 23:57:29,875 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> REGISTER '/home/acadgild/install/pig/pig-0.16.0/lib/piggybank.jar';
grunt> A = load '/home/acadgild/akshat/PIG_SESSION/AirportUseCase/DelayedFlights.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',','NO_MULTILINE','UNIX','SKIP_INPUT_HEADER');
2018-09-17 23:57:29,967 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-09-17 23:57:29,967 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> B = FOREACH A GENERATE (chararray)$17 as origin, (chararray)$18 as dest, (int)$24 as diversion;
grunt> C = FILTER B BY (origin is not null) AND (dest is not null) AND (diversion == 1);
```

Output:



```
2018-09-17 23:57:43,856 [main] INFO  org.apa
((ORD,LGA),39)
((DAL,HOU),35)
((DFW,LGA),33)
((ATL,LGA),32)
((ORD,SNA),31)
((SLC,SUN),31)
((MIA,LGA),31)
((BUR,JFK),29)
((HRL,HOU),28)
((BUR,DFW),25)
grunt>
```

Script explanation:

→registering the piggybank jar to use the CSVExcelStorage class.
→loading the dataset using CSVExcelStorage.

→ using foreach getting origin, destination and diversion column details.
→filtering and keeping rows that have origin and destination not null and diverted as 1.
→grouping filtered data by origin and destination.

→generating grouped column(origin and destiantion) and count of deviation.
→ordering in descending as per dept delay

→Limiting result to top 10.